Protein Structure and Visualization - Answers

By Anne Mølgaard and Thomas Blicher

Q1 *Rhamnogalacturonan acetylesterase in UniProt:*

a) The signal peptide is from residue number 1 to 17.
b) The mature protein is from residue number 18 to 250, which means that the protein consists of 233 residues.
c) The active site is made up of three residues. The first is Ser26, the two others are Asp209 and His212.
d) The protein is post-translationally modified, having two sites of N-glycosylation at121 and 191.

Q2 *Are all hits are relevant if you are looking for a representative structure of the sequence you entered? Which parameters should you look at to make this decision?*

As we have performed a sequence search, use only parameters relating to alignment to judge which hits are relevant. Although you could have a look at the alignment score, alignment length and number of identical, aligned residues, all of that information is combined in the Expect value (E-value). Lower E-values are better, as they indicate the number of hits one would "expect" to find by chance with the given alignment score when searching the database. It turns out that there are currently only five structures (February 2012), which are good representatives of RGAE: 1deo, 1dex, 1k7c, 1pp4 and 3c1u. Other structure hits (1pvx and 1fa2) only cover a small part of the input sequence and have very low sequence similarity to our search sequence; they are therefore not representative of RGAE. **Remember**: Depending on what you want to do with your structure hit, other factors not necessarily directly related to sequence may affect your choice of representative structure. Such factors include structure quality, presence/absence of particular ligands, protein conformation or specific experimental conditions. Thus, you may occasionally choose to work with structures of homologs or mutant proteins, if they contain the information (e.g. ligands) or have the quality you need.

Q3 *Choose the best structure that has sulfate ions bound. Which one did you choose? Why?*

Both 1deo and 1k7c have sulfate ions bound. However, the resolution of 1k7c is 1.12 Å, which is better than the 1.55 Å resolution of 1deo (**Note**: 1.55 Å is very good under most circumstances and resolutions better than this – i.e. lower – are not common). The R_{free} is 0.134 for 1k7c, which is also better than 0.200 for 1deo (again this is mostly a function of resolution although other factors such as data quality and refinement protocols also contribute).

Q4 *What is the residue name for the sulfate ions?*

SO4

Q5 *Click on H(ide) and select "waters". What happened?*

When water molecules are hidden with the Hide command (button), they will simply be switched off. To turn them on again simply click on S(how) – nonbonded.

Q6 *The active site of RGAE.*

The active site residues are: Ser9, His195 and Asp192. These numbers do not directly correspond to the information in Swiss-Prot entry. The reason is that residue numbering in the PDB file starts with the residues of the native protein, i.e. the mature protein sequence without the signal peptide. This means that all residue numbers in the PDB file are off by 17.

PyMOL magic:

Instead of manually looking for the catalytic site residues in the active site, you could use the following two commands in PyMOL:

```
select triad_his, (byres resn his within 3 of resn asp) and
(byres resn his within 3 of resn ser)
```

This first command selects all those histidine residues (selection name: triad_his), which are simultaneously within 3 Å of any aspartic acid residue **AND** within 3 Å of any serine residue, which is the case for histidine residues found in catalytic triads of the kind we are looking for. The byres modifier ensures that we select entire residues and not just those atoms, which fulfill the distance requirement. It turns out that there is only one such residue in the 1k7c structure. To select the other residues of the catalytic triad, simply write:

```
select triad, triad_his or (byres resn asp within 3 of
triad_his) or (byres resn ser within 3 of triad_his)
```

This second command selects the histidine residue found with the first command (again) along with aspartic acid and serine residues within 3 Å of that histidine. And voilà, you have found your catalytic triad ☺

To get the residue names and numbers, either click the residues in the viewer window or type the following three commands:

```
triad_list = []
iterate triad and name CA, triad_list.append((resn, resi))
for pair in triad_list: print pair[0], pair[1]
```

This should print the following information in the command window:

```
SER 9
ASP 192
HIS 195
```

Note: In the general case, you can modify the commands above to look for other kinds of arrangements of residues, but you will of course need to know rather accurately how the residues of interest are positioned relative to each other.